2020 IEEE International Conference on Sustainable Engineering and Creative Computing (ICSECC) | 978-1-7281-7588-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSECC51444.2020.9557529

Siamese Network's Performance for Face Recognition

Steven
Faculty of Information and Technology
Tarumanagara University
Jakarta, Indonesia
steven.535160001@stu.untar.ac.id

Janson Hendryli
Faculty of Information and Technology
Tarumanagara University
Jakarta, Indonesia
jansonh@fti.untar.ac.id

Dyah Erny Herwindiati
Faculty of Information and Technology
Tarumanagara University
Jakarta, Indonesia
dyahh@fti.untar.ac.id

***Abstract*—Performance of Siamese network for real-time face recognition software in a one-shot learning setting is discussed in the paper. Two loss functions for the Siamese network are also compared, which are the contrastive loss and the triplet loss. Initially, a multitask cascaded neural network detects faces from a webcam, and the Siamese network matches the detected faces to the user's registered face. In the experiment evaluation, we find that the Siamese network with contrastive loss achieves better performance. The accuracy is 0.8875. However, the model with triplet loss has an accuracy of0.85.**

Keywords—real-time face recognition, siamese network, convolutional neural network, contrastive loss, triplet loss.

I. INTRODUCTION

The application of biometric security system is widely adopted. Related to the aforementioned example, a face recognition system for recording employee attendance cannot be trained with the whole employee's face dataset. It is impractical to retrain the model every time a new employee joins the company. The system has to take into account of an employee who is leaving the company. Retraining a model takes a lot of time and computational powers. One of the implementations that have already existed in the Baidu attendance system. The system used a deep learning model trained with Siamese Network to identify its employee attendance. It identified the employee's face taken from the camera and matched it with the employee's images registered in the system.

The advancement of deep learning models popularizes the convolutional neural networks (CNN) as reliable classification models for image data. CNN is accurate for image classification tasks due to the huge amount of data available in the training stage [1]. However, if there are new classes or labels introduced to the classification task, the model must learn again, either from the beginning or utilize the transfer learning method [3]. For example, we have to train the CNN model to every person in the face recognition task where the model has to classify the person the face belongs to.

Though the underlying architecture uses CNN, The Siamese network is not the same as conventional CNN, where it is trained to classify images. The Siamese network is trained to identify or differentiate between pair of images. One of the essential differences between these architectures is that conventional CNN could only identify images that belong to a class is in training sets. Otherwise, the model could not identify them. On the other hand, the Siamese network could identify images regardless of which class they belong to. This particular architecture is being trained to differentiate one image from others.

This paper explores a real-time face recognition system where the model does not need to be retrained to classify a new person. The system works in real-time to detect and recognize the faces of the video feed. The model is called the Siamese network. The experiments are initiated by some works that have been done recently. The work that is proposed by Omkar et al., 2015, which focused on developing a face recognition model using Siamese Network with Triplet Loss. In this work, every subject's image has six images, consisting of only the subject's face. The training sets they used were from the CelebFaces dataset and WDRef. The model was evaluated using Labeled Face in Wild (LFW) with an accuracy of 98.95% [1]. Bella et al., 2020, proposed another work that focused on developing a voice authentication model using Siamese Network with Binary Cross Entropy. The model was evaluated with an F1-Score of 48 [2].

The CNN is trained to extract the distinguishing feature in human faces by feeding the model a different and same person. The model used two identical CNN architectures and Euclidean Distance to measure similarity distance between inputs. The architecture is trained with contrastive loss and triplet loss functions known for their performance in distance metric learning. We used the AT&T Faces dataset and images taken with Webcam Logitech C270 for training sets. The model was evaluated with 101 subjects who are seated 20 centimeters apart from the webcam. The goal of this paper is to see the performance of the Siamese network for real-time face recognition using AT&T Faces dataset and images taken with Webcam Logitech C270.

II. METHODS

Our real-time face recognition system consists of two stages. We employ the multitask cascaded neural network to detect faces from a video feed in the first stage. The system uses the Siamese network to determine the individual's identity by matching the detected face to registered users' faces in the database. We first describe the convolutional

978-1-7281-7588-1/20/$31.00 ©2020 IEEE

neural network, which is the underlying method in the multitask cascaded neural network and the Siamese network.

detection result and its positions. The overall illustrations of the stages are shown in Fig. 1

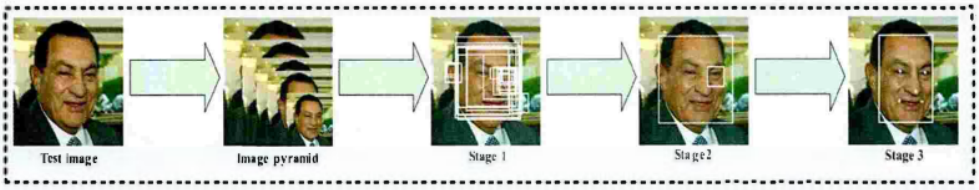

Fig. 1. Three steps of the MTCNN to detect human face from an image [5]

A. Convolutional neural network

The convolutional neural network, frequently abbreviated as CNN, is one of the deep learning methods that are popular in image classification tasks. Its architecture consists of a neuron that has weight and bias [1], which interconnects to neurons in the previous layer. There are three main layers in the CNN architecture: convolutional layer, pooling layer, and fully connected layer [1]. The convolutional layer can be assumed as the main part of a CNN model. This layer has filters updated in every training iteration that aims to learn the input image's best representation corresponding to the output. The convolutional operation in this layer corresponds to a dot product of the filter and input matrix. The convolution of an image I with the size of H × W and a filter K of size k1 × k2 is as in Eq. 1.

$$(I*K)_{ij} = \sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}\sum_{c=1}^{C} K(m,n)\cdot I_{i+m,j+n} + b \quad (1)$$

There are three hyperparameters that affect the output size of the convolution layers: depth, stride, and zero paddings. The depth of the output corresponds to the number of filters. Stride is the size of the leap of the filter when doing the dot product operation. Meanwhile, zero-padding adds zero value around the input matrix. The pooling layer is usually used after the convolution layer and aims to reduce the output size and the computation cost. Current research usually employs a dropout strategy to avoid overfitting when training CNN. Let a CNN model with 10 neurons in a layer, a dropout probability set to 0.2 means the model selects two random neurons to be taken out during the training process [4].

B. Multitask cascaded neural network

Multitask cascaded neural network (MTCNN) introduced by [5] is a neural network specifically trained to detect faces in an image. It is trained using the Annotated Facial Landmarks in the Wild (AFLW) dataset and evaluated using Face Detection Dataset, and Benchmark (FDDB) and Annotated Faces in the Wild (AFW) dataset. The underlying framework of MTCNN uses CNN to detect faces in three stages. The first stage uses shallow CNN to produce candidate windows, refined by a more complex CNN in the next stage. The final stage uses more powerful CNN to output the

C. Siamese Network

Siamese network is an architecture used to do one-shot learning [6]. In this work, the Siamese network receives the detected faces from MTCNN and uses them to recognize the individual. There are two types of Siamese network differed by the number of inputs used by the model. The first type is a Siamese network with three input images and uses triplet loss [7] as the loss function. The inputs are one anchor image, one positive image with the same class as the anchor image, and one negative image [8]. The triplet loss can be defined as in Eq. 2.

$$L = \max(d(a,p) - d(a,n) + margin, 0) \quad (2)$$

where d(,) denotes a distance metric, a is the anchor image, p is the positive image, and n denotes the negative image. The distance metric can be a Euclidean distance or a Mahalanobis distance.

Another type of Siamese network uses contrastive loss as the loss function and a pair of images as its input. The goal of this model is to be able to differentiate whether two inputs belong to the same class or not [9]. The contrastive loss function is defined as in Eq. 3.

$$L = (1-Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}\{\max(0, m - D_w)\}^2 \quad (3)$$

Where Y denotes the output label, D_w is the distance function between a pair of samples, and $m > 0$ denotes a margin. The output of this type of network is in a distance value. The Euclidean or Mahalanobis distance can calculate the distance between encoded images 1 and 2 experiments. The experiment is conducted using the AT&T faces dataset, which consists of 10 different images of each of 40 distinct subjects with varying lighting, facial expressions, and details [10]. Figure 2 shows the example of faces in the dataset. The images in the dataset are all taken with a dark homogeneous background.

Furthermore, we also collect our own faces dataset from 50 different subjects of 5 different images each. Moreover, the images are all in 96*96 pixels in grayscale format and have been normalized. We use two dataset scenarios to train, validate, and test the model. The training and validation process uses the AT&T faces dataset, which is split 80% for training and the remainder 20% for validation. Meanwhile, to test the model learned from the training process, we use our

collection of faces captured from a Logitech C270HD webcam.

TABLE I. THE RESULT OF THE EXPERIMENTS SHOWING THE MODEL WITH THE BEST ACCURACY AND ITS CORRESPONDING EER AND THRESHOLD

Loss	Accuracy			EER	Threshold
	Train	*Val*	*Test*		
Triplet	0.99	0.99	0.85	0.05	0.634
Contrastive	0.9965	0.9889	0.8875	0.025	0.302

For the model, we experiment on two types of Siamese network architectures, which are Siamese with triplet loss and contrastive loss. Every training process also uses an early stopping procedure where the training is stopped when the loss after some epochs decreases. We try different epochs and learning rates for each architecture to find the best model.

To implement real-time face detection using the learned model, a threshold to accept or reject the recognition result is calculated using the equal error rate (EER).

Fig. 3. The AT&T dataset consists of faces from 40 different individuals. (Source: https://medium.com/@MarynaL./eigenfaces-

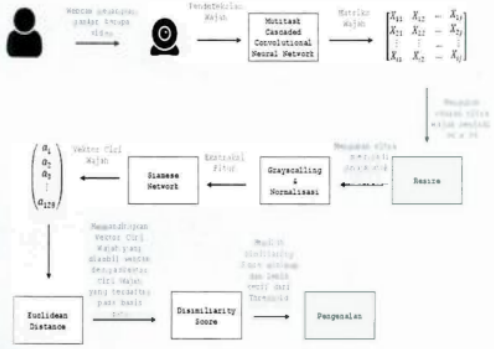

Fig. 5. Recognition Schema

The threshold value is computed by finding the intersection of the false acceptance rate (FAR) with the false rejection rate (FRR). The FAR indicates a probability of someone who is not registered in a face recognition system but recognized as one of the subjects in the system.

In contrast, the probability of registered users not recognized by the system is indicated by the FRR. Thus, the

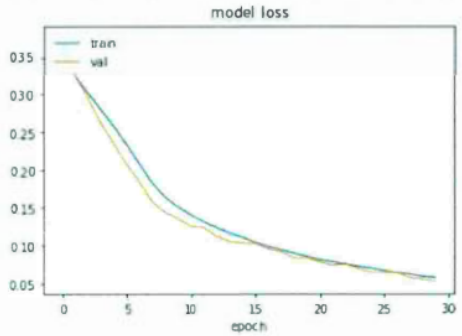

Fig. 2. The training and validation loss of the Siamese network model trained using contrastive loss.

FAR and FRR are calculated by dividing the number of false acceptance and false rejection, respectively, over the number of recognition attempts.

Finally, after obtaining the threshold, the system is tested with over 80 face images of 16 subjects in the test set, of which 8 subjects are registered in the system, while the rest are unregistered. Some testing scenarios include faces in a straight view and tilted either to the left or right (maximum 30 degrees). All faces are 45cm away from the webcam.

III. RESULTS

The result of the experiment is shown in Table I. Overall, we can say that the model with contrastive loss (CL) achieves better performance than the model using triplet loss (TL) with 0.8875 in test accuracy. Fig. 4 and 5 show the training and validation loss of both models. Although the model is trained in 50 epochs, the early stopping mechanism halts the training sooner because the loss is not decreasing after 3 iterations.

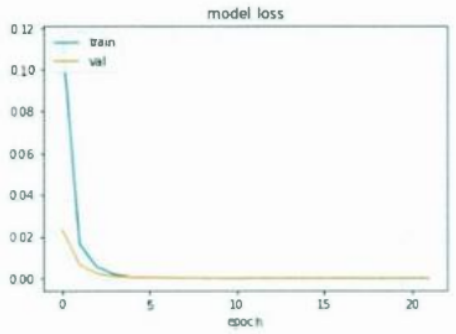

Fig. 4. The training and validation loss of the Siamese network model trained using triplet loss

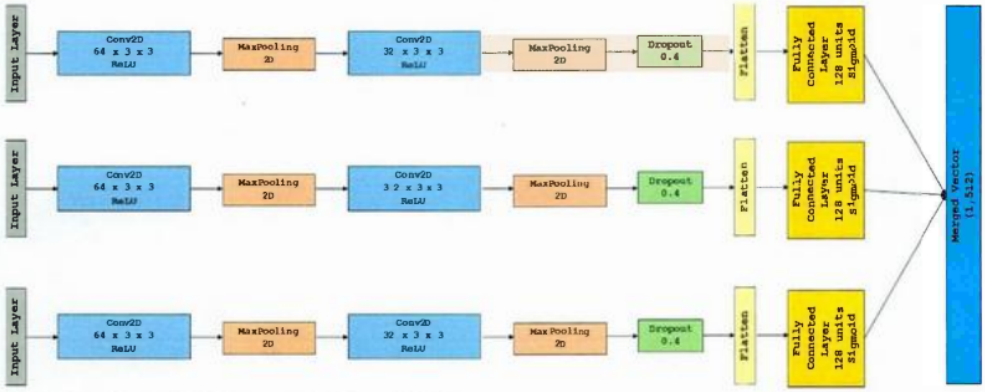

Fig. 7. The architecture of the Siamese network with contrastive loss.

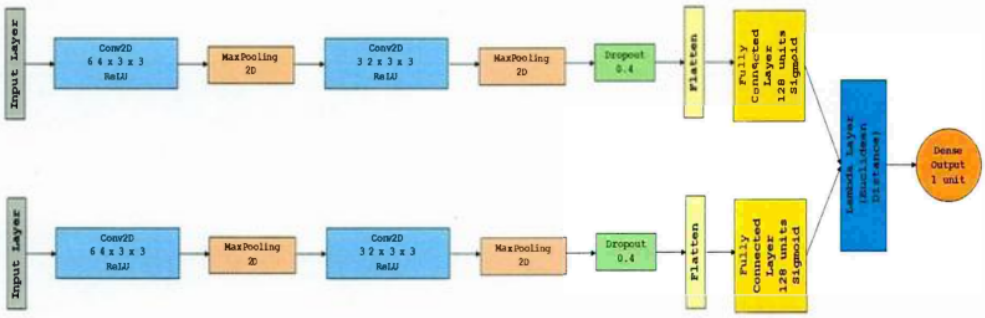

Fig. 8. The architecture of the Siamese network with triplet loss.

Fig. 8 shows the illustration of threshold calculation for the model with triplet loss. The value where the FAR and FRR intersects constitutes the threshold and EER, which are 0.6340 and 0.05, respectively. If the output of the Siamese network, which is the distance among the inputs, is larger than the threshold, the model accepts or matches the person's face to a registered user's face. The lower EER also shows a higher accuracy of the model.

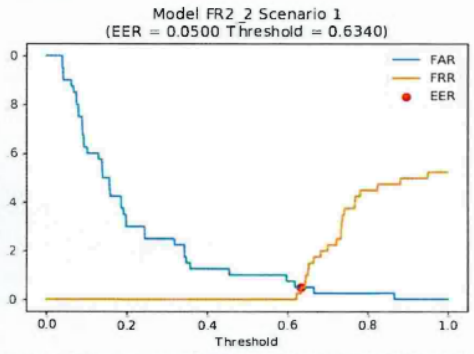

Fig. 6. The equal error rate (EER) calculation to find the best threshold

Both models are trained in 50 epochs and use RMSProp optimizers. The CNN architecture consists of 2 convolutional layers (64 hidden units for the first layer, followed by a layer of 32 hidden units) with kernel size 3x3. Each convolutional layer is immediately followed by 2D max pooling. Following the convolutional layers is the fully connected layer, consisting of 128 sigmoid units in the CL model and a merged layer in the TL model. The dropout regularization of 0.4 is also used in both training process. Finally, the CL model is trained using a batch size of 128, while the TL model uses a batch size of 256. The learning rate for the best CL model training is 0.001 and 0.0001 for the best TL model. See Fig. 6 and 7 for the illustration of the Siamese network architecture.

IV. CONCLUSION

In this work, we demonstrate the Siamese network model for a real-time face recognition system. It matches the face from a webcam to registered faces in the database by initially learning the human face's representation using convolutional neural networks and measuring the distance. Using equal error rate calculation, the threshold to accept or reject the recognition attempt is established. In addition to the AT&T faces dataset, we also collect faces from several subjects in various settings for the model training and testing process. We experiment with two types of losses, which are

contrastive loss and triplet loss, to train the Siamese network. The experiment finds the contrastive loss has better overall performance than the triplet loss model. The equal error rate also shows that the model has high accuracy (lower equal error rate). A few things to consider for future works are experimenting on more real-time variations in lighting, facial expression, and details. Furthermore, using a larger dataset with more face images for training and validation can be considered to better understand the architecture's performance. Ultimately, the system should be able to recognize the faces quickly and with reliable performance.

ACKNOWLEDGMENT

We gratefully acknowledge Titan X and Titan Xp GPUs' donation from the NVIDIA Corporation used in this paper to train the model.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 26th British Machine Vision Conference, vol. 1, no. 3, 2015.

[2] Bella, J. Hendryli, and D. E. Herwindiati," Voice Authentication Model for One-time Password Using Deep Learning Models", Proceedings of the 2020 2nd international Conference on Big Data Engineering and Technology , 2020., pp. 35-39.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," Synthesis Lectures on Computer Vision, vol. 8, no. 1, 2018.

[5] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[6] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in ICML deep learning workshop, vol. 2. Lille, 2015.

[7] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 459–474.

[8] C. Francois, Deep learning with Python. Manning Publications Company, 2017.

[9] J. Loy, Neural Network Projects with Python: The ultimate guide to using Python to explore the true power of neural networks through six projects. Packt Publishing, 2019.

[10] AT&T Laboratories Cambridge, "The Database of Faces," https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (accessed Aug. 19, 2019).

DYAH E H 'SIAMESE NETWORK'S' PROSIDING_0001

ORIGINALITY REPORT

6%	3%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

1%

★ export.arxiv.org

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		